DOCUMENT RESUME

ED 437 399 TM 030 551

AUTHOR Onwuegbuzie, Anthony J.; Daniel, Larry G.
TITLE Uses and Misuses of the Correlation Coefficient.
PUB DATE 1999-11-00
NOTE 58p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Point Clear, AL, November 17-19, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Causal Models; *Correlation; Effect Size; *Error of Measurement; *Research Methodology
IDENTIFIERS Bootstrap Methods; Jackknifing Technique

ABSTRACT

The purpose of this paper is to provide an in-depth critical analysis of the use and misuse of correlation coefficients. Various analytical and interpretational misconceptions are reviewed, beginning with the egregious assumption that correlational statistics may be useful in inferring causality. Additional misconceptions, stemming from researchers' failure to recognize that correlation coefficients are specific cases of the general linear model (GLM), and, therefore, bounded by GLM assumptions, are also discussed. Other inappropriate practices are highlighted, including failure to: (1) consider the statistical assumptions underlying correlation coefficients; (2) interpret confidence intervals and effect sizes of correlation coefficients; (3) interpret p-calculated values in light of familywise Type 1 error; (4) consider the power of tests of hypotheses; (5) consider whether outliers are inherent in the data set; (6) recognize how measurement error can affect correlation coefficients; and (7) evaluate empirically the replicability of correlation coefficients. A heuristic example is used to illustrate how jackknife and bootstrap methods can identify unstable correlation coefficients derived from a given sample. (Contains 2 tables, 1 figure, and 80 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 437 399

Running Head: USES AND MISUSES OF THE CORRELATION COEFFICIENT

Uses and Misuses of the Correlation Coefficient

Anthony J. Onwuegbuzie

Valdosta State University

Larry G. Daniel

University of North Texas

TM030551

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

A. Onwuegbuzie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Paper presented at the annual meeting of the Mid-South Educational Educational Research Association (MSERA), Point Clear, Alabama, November 17-19, 1999.

BEST COPY AVAILABLE

Abstract

The purpose of this paper is to provide an in-depth critical analysis of the use and misuse of correlation coefficients. Various analytical and interpretational misconceptions are overviewed, beginning with the egregious assumption that correlational statistics may be useful in inferring causality. Additional misconceptions, stemming from researchers' failure to recognize that correlation coefficients are specific cases of the general linear model (GLM), and are therefore bounded by GLM assumptions, are also discussed. Other inappropriate practices are highlighted, including failure to consider the statistical assumptions underlying correlation coefficients, failure to interpret confidence intervals and effect sizes of correlation coefficients, failure to interpret *p*-calculated values in light of familywise Type I error, failure to consider the power of tests of hypotheses, failure to consider whether outliers are inherent in the data set, failure to recognize how measurement error can affect correlation coefficients, and failure to evaluate empirically the replicability of correlation coefficients (i.e., internal replication). A heuristic example is utilized to illustrate how jackknife and bootstrap methods can identify unstable correlation coefficients derived from a given sample.

Uses and Misuses of the Correlation Coefficient

In 1889, Francis Galton published his book entitled, *Natural Inheritance* (as cited in Gigerenzer et al., 1989), in which he introduced the concept of correlation and established the beginning of modern statistics (Coblick, Halpin, & Halpin, 1998). Building on Galton's work, Karl Pearson extended the concepts of correlation and the normal curve, developing the product-moment correlation coefficient and other types of correlation coefficients (Coblick et al., 1998). It is likely that neither Galton nor Pearson envisioned the impact that the correlation coefficient would have in the field of research, particularly within the social and behavioral sciences. Indeed, today, slightly more than a century later, the correlation coefficient is undoubtedly the most common statistic used in research involving inferential analyses.

Correlation coefficients in general and the Pearson product-moment correlation coefficient in particular, are utilized in the majority of studies in educational and psychological research, either as a primary mode of analysis in which major hypotheses are tested, or as part of a secondary analysis, providing background information regarding relationships among variables of interest prior to or following a more complex statistical analysis. Unfortunately, although the use of the correlation coefficient is justified in many situations, like all statistical indices, it is subject to misuse. That is, many examples exist wherein this statistic is misinterpreted.

Purpose of the Present Paper

The purpose of the present paper is to provide an in-depth critical analysis of the use of correlation coefficients. As part of this critique, a variety of analytical and

interpretational misconceptions are presented. Phenomena such as the *crud factor* and *positive manifold* are discussed. Perhaps the most serious error in interpreting correlational analyses is the failure to evaluate empirically the replicability of the coefficients obtained with a given sample (Thompson, 1999). Although, as noted by Thompson (1999), the most valid and appropriate manner of assessing the empirical replicability of findings is unequivocally via new and independent samples (i.e., external replication; Huberty & Wisenbaker, 1992), a paucity of researchers are able or willing to conduct external replication analyses. As a more feasible alternative, analyzing data from the sample at hand (i.e., internal replication; Huberty & Wisenbaker, 1992) is recommended (Thompson, 1994a). Unfortunately, few researchers conduct internal replications either. Thus, a heuristic example will be utilized to illustrate how two internal replication techniques (i.e, jackknife and bootstrap methods) can identify unstable correlation coefficients derived from the full sample.

Most of the discussion below is not new to the literature or even to our own writing (e.g., Daniel, 1989, 1998a, 1998b; Onwuegbuzie, 1999a). However, the fact that many beginning researchers misuse and misinterpret correlation coefficients (perhaps as a result of the "mythology of statistics"–Daniel, 1997; Kerlinger, 1960), as well as the fact that very few of even the most experienced researchers conduct internal replications as part of their data analyses justify our attention to this topic. In any case, most methodological papers in this area have dealt with issues in piecemeal fashion, whereas the current paper attempts to provide a more comprehensive discussion of issues for consideration when utilizing correlation coefficients. Moreover,

most quantitative methodologists tend to focus on issues pertaining to more complex inferential techniques such as multivariate analyses. As such, methodological and conceptual issues concerning the correlation coefficient recently have received scant focus. Yet, common errors surrounding this seemingly simplistic measure of relationship continue to permeate the literature.

Overview of the Pearson Product-Moment Correlation Coefficient

As all researchers know, correlation coefficients, which can vary from -1 to +1, help to determine both the magnitude and direction of pairwise variable relationships. The sign of the coefficient tells us whether the relationship is positive or negative, whereas the numerical part of the coefficient indicates the magnitude of the correlation. The closer the correlation coefficient is to 1 or -1, the greater the relationship between the variables.

There are various "zero-order" correlational statistics (i.e., correlational measures of bivariate relationships that do not include adjustments for other variables), including the Pearson product-moment correlation coefficient (r), Spearman's rho (ρ), Kendall's tau (τ), point biserial correlation(r_{pb}), biserial correlation, phi (ϕ), and tetrachoric correlation. Pearson's r may appropriately be considered for use when both variables represent either interval or ratio scales of measurement. Spearman's rho is most appropriate when both variables represent the ordinal scale of measurement. Kendall's tau is similar to Spearman's rho inasmuch as it is suitable for use when the variables are in the form of ranks. The major difference between Kendall's tau and Spearman's rho is the former tends to be used when tied ranks are present. Point biserial

coefficients are appropriate when one variable is measured on the interval or ratio scale and the other variable is a dichotomous variable which takes values of 0s and 1s (e.g., scores for items on a multiple-choice test). A point biserial correlation coefficient is a special case of the Pearson product-moment correlation coefficient, and it is computationally a variant of the *t*-test . The biserial correlation coefficient is similar to the point biserial coefficient, except dichotomous variables are artificially created (i.e., using "cutoff" values of a previously continuous variable). Phi coefficients are used when both variables represent true dichotomies. A phi coefficient, like the point biserial correlation, is directly derived from the Pearson product-moment correlation. Tetrachoric correlations are utilized when each variable is created through dichotomizing an underlying normal distribution. (Various "higher order" correlations, such as partial correlations, semipartial [part] correlations, and multiple R, are also frequently used in the social sciences; however, these correlational statistics are considered beyond the scope of the present paper.)

In the remainder of this paper, methodological and conceptual errors concerning the most frequently utilized type of zero-order correlation coefficient, namely, the Pearson product-moment correlation coefficient, will be presented. The Pearson *r* was chosen as the focus of the paper in that most readers will be familiar with *r* even though knowledge of various other zero-order bivariate measures may be somewhat more limited. With a few exceptions, all other empirical measures of bivariate relationships are susceptible to these same flaws.

Flaw 1: Inadequate Checking of Statistical Assumptions

Many misconceptions stem from a failure to recognize that correlation coefficients are a specific case of the general linear model (GLM). That is, correlation coefficients are special cases of all other families of the GLM, including *t*-tests, multiple regression, analysis of variance, canonical correlation, and structural equation models (Cohen, 1968; Knapp, 1978; Thompson, 1998a). Indeed, Pearson product-moment correlation analysis is directly analogous to simple linear regression (Myers, 1986). As such, many of the assumptions which apply to these more complex members of the GLM, also are pertinent to correlation coefficients.

The major assumptions for conducting a null hypothesis significance test (NHST) for the Pearson's product-moment correlation coefficient are the same as that for simple linear regression. Specifically:

1. Each observation of the dependent variable (Y) must be statistically independent of every other observation.
2. The dependent variable (Y) must be normally distributed.
3. The variability in scores for the dependent variable is approximately the same at all values of the independent variable (i.e., the "conditional distribution" or "homoscedasticity" assumption).

The first assumption, namely, independence of observations can be assessed by carefully examining the research design. For example, if the dependent variable is a test score, one should check that each student attempted the examination form independently. The second assumption, normality, should be assessed by both

graphical and statistical means. With respect to the former, frequency histograms could be utilized. Expected normal probability plots are even more informative for assessing normality. These plots represent the difference between the observed normal value for each case (i.e., the *z* score that a case has in the observed distribution) and the corresponding expected normal value (i.e., the *z* score that a case with the observed rank holds in the normal distribution), such that if the observed scores are normally distributed, then the bivariate points lie on the diagonal running from the lower left to the upper right of the two-dimensional grid.

In addition to graphical checks of normality, the skewness and kurtosis can be assessed for magnitude by comparing these values to their corresponding standard errors. These four statistics are available in the Statistical Package for the Social Sciences (SPSS; SPSS Inc., 1999). Indeed, a formal test of statistical significance can be conducted by utilizing the fact that the ratio of the skewness and kurtosis coefficients to their respective standard errors are themselves normally distributed. Most other statistical packages print as options skewness and kurtosis coefficients but not their standard errors. However, these standard errors can be approximated manually (the standard error for skewness is approximately equal to the square root of $6/n$, and the standard error for kurtosis is approximately equal to the square toot of $24/n$, where n is the sample size). Large skewness and kurtosis coefficients affect the Type I and Type II error rates. For example, non-normal kurtosis tends to produce an underestimate of the variance of a variable, which, in turn, increases the Type I error rate (Tabachnick & Fidell, 1996). Thus, if the assumption of normality is found to be violated, other

correlational techniques such as Spearman's rho should be considered.

The third assumption, namely, homoscedasticity, could be assessed by examining bivariate scatter plots. When points on this plot appear to take on a "funnel" shape, it is likely that the assumption of homoscedasticity is not met. The more funnel shaped the bivariate points are, the greater the level of heteroscedasticity. In extreme cases of heteroscedasticity, data transformations (Fox, 1997) should be considered. If heteroscedasticity is suspected, then the product-moment coefficient should be abandoned for tests that are designed for unequal variance conditions (Huck & Cormier, 1999).

Interestingly, the above three assumptions are often stated in terms of the errors (ϵs) of the simple linear regression analog to Pearson's product-moment correlation coefficient. That is, the error components of the model must be normal, equally variable, and independent of each other, with a mean of zero and constant variance. The general equation underlying the simple linear regression (SLR) model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 is an additive weight (or constant), β_1 is a multiplicative weight (unstandardized regression coefficient), and ϵ represents the random deviation of an observed y value from the estimated sample regression line (i.e., the model error). Instead of computing the correlation coefficient, an analyst could directly undertake an SLR analysis and then assess the assumptions regarding normality and homoscedasticity by examining the residuals. Typically, either standardized or studentized residuals could be analyzed (Fox, 1997). (The absolute value of the

product-moment correlation coefficient can be derived by taking R, the square root of the coefficient of determination, R^2, from the SLR model; however, because the variable weighting procedures used in linear regression always result in a positive value of R [$0 \leq R \leq 1$], the sign of the regression coefficient must also be consulted in order to determine the directionality of the relationship.)

In examining residuals, one also should assess outliers. These observations are extreme values which exert undue influence on models, and thus lead to both Type I and Type II errors. Unfortunately, it is typically unclear as to what effect that outliers have in a particular analysis (Tabachnick & Fidell, 1996). Nevertheless, by identifying potential outliers, analysts can make a decision as to whether to delete or retain the case, to present results both with and without the outlying points included, or to transform one or both of the variables involved. Popularly-used statistical packages, such as SPSS include a variety of helpful statistics for detecting outliers. Alternatively, in the presence of outliers one might use more robust correlations such as Spearman's rho, Kendall's tau, and "percentage-bend" correlation (see Wilcox, 1997 for a discussion of robust correlations).

Because a correlation coefficient involves one independent variable, or at least one arbitrarily independent variable (in cases when the temporal order between the two variables, such as self-esteem and anxiety, is unclear), multicollinearity is not an issue when examining bivariate relationships. However, there are two more assumptions that are often not checked, and yet, if violated, could invalidate the NHST for Pearson's product-moment correlation coefficient. These assumptions are

4. All variables are measured without error;

and 5. The relationship between the two variables is linear.

When conducting statistical significance tests of correlation coefficients, as with all NHSTs, it is hoped that all variables are measured with little or no error. One way of assessing error of measurement is through reliability coefficients. Yet, surprisingly, relatively few researchers report reliability coefficients as they pertain to their sample (Eason & Daniel, 1989; Meier & Davis, 1990; Onwuegbuzie, 1999a; Thompson, 1998a, 1999; Willson, 1980), often because of a failure to realize that reliability and validity are a function of scores, not of instruments (Pedhazur & Schmelkin, 1991). Using measures with low reliability typically increases Type II error (i.e., reduces statistical power) by attenuating relationships. However, without information about the reliability of the scores on each of the measures, it is not possible to assess the extent to which the NHST for the product-moment correlation coefficient is affected. However, if reliability estimates for the scores on the two variables are computed, the analyst will be able to determine not only how much error is in each set of scores, but also the maximum correlation one might expect under the conditions of the two reported reliability estimates, considering that "the correlation between scores from two tests cannot exceed the square root of the product for reliability [of scores] in each test" (Locke, Spirduso, & Silverman, 1987, p. 28). For example, if the reliability estimate for scores on variable *X* is .80, and the reliability estimate for scores on variable *Y* is .60, the correlation between the two variables cannot exceed |.69|). Thus, as recommended by Thompson (1998a, 1999), researchers *always* should report reliability coefficients for

their own data.

The assumption of linearity means that there is a straight line relationship between the two variables of interest. This is a crucial assumption because Pearson's product-moment correlation coefficient only describes the linear relationship between variables; any non-linear relationship that exists is not captured by Pearson's *r*. Non-linearity (i.e., curvilinearity) can be examined either via bivariate scatterplots or from residual plots stemming from SLR analyses. As obvious as the importance of linearity is, and as easy as it is to check this assumption, it appears that few researchers routinely do so. Yet, there are many variables, for example, variables that are a function of time, which are susceptible to non-linearity. Thus, it should not be assumed that all bivariate relationships of interest are linear. Indeed, as noted by Maxwell and Delaney (1990, p. 361), "many graduate students have been embarrassed by writing theses based on computer-generated summary statistics, only later to learn that the results were nonsensical."

In sum, providing that the assumptions are met, the use of Pearson's product-moment correlation coefficient is justified. However, when one or more of the assumptions discussed above are grossly violated, the ensuing product-moment correlation coefficient may be invalid and, more importantly, any interpretation of it will be misleading. Thus, it is essential that researchers not only routinely assess all pertinent validity assumptions associated with the Pearson *r*, but they also should make reference to the results of these checks in their reports. Indeed, journal editors should strongly encourage this practice, as such information often can be disseminated in a

few sentences. Currently, a paucity of researchers discuss validity assumptions in their articles (Keselman et al., 1998).

Flaw 2: Failure to Adjust for Type I Error When Conducting Multiple Tests

When the statistical significance of more than one correlation coefficient is tested within a study, as is typically the case when bivariate relationships are of interest, adjustments for inflated Type I error rates should be made. For example, when a researcher conducts a statistical significance test for only one *r* coefficient, the probability of rejecting a true null hypothesis is equal to the critical *p* (α)value for that test. However, as Stevens (1996, pp. 6-9) illustrated, when *k* correlational tests are tested for statistical significance, assuming independence of each test, with testwise error probabilities of $\alpha_1, \alpha_2, \ldots, \alpha_k$ for the *k* tests, the *overall alpha* (or "familywise" alpha) level will actually exceed the value of any of the individual testwise alphas:

$$\text{overall } \alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)\ldots(1 - \alpha_k)$$

In the case that all *k* statistical significance tests employ the same testwise alpha level (α_T), the equation could be simplified:

$$\text{overall } \alpha = 1 - (1 - \alpha_T)^k$$

Using this latter formula, in the case of computing and testing the statistical significance of 10 independent *r*'s from a given sample at the .05 level of probability, the overall Type I error probability would actually be not a meager 5%, but a whopping 40.1%! However, because the 10 correlation coefficients computed from the same sample would likely not be completely independent (due to intercorrelations of the variables as a set), the overall Type I error probability rate would actually fall

somewhere between the testwise alpha (5%) and the overall alpha (40.1%), an "improvement" of the odds that should give the researcher, at best, only minimal comfort.

In handling this problem, all the tests of statistical significance can be undertaken either by using the same "adjusted" alpha level via techniques such as the Bonferroni adjustment (Tabachnick & Fidell, 1996), or by making some tests more liberal than others in the set of correlations analyzed, via methods such as the Holms procedure (Huck & Cormier, 1999). Whatever technique is used, it is important to attempt to ensure that the actual Type I actual error does not exceed its nominal value. Unfortunately, many researchers do not make adjustments for Type I error when conducting multiple tests of correlation coefficients (Onwuegbuzie, 1999a). Further, in the event that the variables being correlated represent two discrete sets (i.e., predictor and criterion sets), it would behoove the researcher to utilize a more advanced correlational procedure (e.g., multiple regression, canonical correlation). Use of more advanced procedures serves to minimize the number of statistical significance tests employed, thereby defeating the Type I error inflation problem, in addition to representing a more realistic picture of the multivariable reality in which the variables actually occur (Fish, 1988; Stevens, 1996).

Flaw 3: Failure to Consider the Power of Tests of Hypotheses

Statistical power is the hypothetical conditional probability of rejecting the null hypothesis (e.g., of no relationship between two constructs) under some alternative hypothesis for the population parameter's value (e.g., a non-zero relationship between

two constructs). Power is affected by three factors: (a) the size of the statistical significance level (increasing alpha increases power but also increases Type I error); (b) the sample size (increasing sample size has the effect of reducing the standard error which, in turn, increases power); and (c) the effect size--the discrepancy between the value of the parameter under the null hypothesis and the value of the parameter under the alternative hypothesis (the larger the difference, the greater the power to detect a difference regarded as notable).

When sample sizes are relatively small, a correlation coefficient that appears to be large may end up being statistically non-significant due to inadequate statistical power. Thus, where possible, researchers should pay attention to sample size prior to collecting data. Although power typically is difficult to calculate for more complex members of the general linear model family, determination of power for tests of correlation coefficients is relatively straightforward. For example, using Table 3.3.5 on pages 92-93 of Cohen's (1988) book, it can be seen that, in order to test a nil null hypothesis (i.e., a hypothesis of zero correlation; Cohen, 1994) for a Pearson product-moment correlation, using an alpha of .05 and a power of .80 (which is deemed to be a desirable combination), a sample size of 28 is needed to detect a large correlation (i.e., $r = .5$), a sample size of 84 is needed to detect a moderate correlation (i.e., $r = .3$), and a sample size of 800 is needed to detect a small correlation (i.e., $r = .1$). When the intention is to test multiple Pearson correlation coefficients, the Bonferroni adjustment should be applied before Cohen's (1988) tables are utilized.

In cases when the researcher has little or no control over the size of the sample

(as appears to be the norm in educational research), *post hoc* power analyses should be undertaken. That is, the resultant sample size should be used to determine the power of the test for an observed correlation coefficient and nominal level of significance. Such post hoc analyses can help analysts to put their findings in the proper statistical context. Yet, few researchers conduct either planned or post-hoc power analyses (Keselman et al., 1998; Onwuegbuzie, 1999a), despite the fact that the power of most studies is unacceptably low (Cohen, 1962, 1965, 1988, 1994, 1997; Schmidt, 1996; Sedlmeier & Gigerenzer, 1989).

Finally, although many recommend that correlation matrices be used to present the results of a correlational study in which there are three or more variables (in order to allow researchers to replicate or to re-analyze their existing data), researchers should refrain from highlighting *all* correlations that are statistically significant at the .05, .01, and .001 levels, as is the current practice--especially when the number of variables included in the table is large. Rather, researchers only should highlight correlations that are statistically significant *after* adjusting for Type I error. In fact, bivariate correlations should be tested for statistical significance only if bivariate hypotheses are of interest. If more complex variable relationships are reflected in a study's hypotheses, and, hence, a bivariate correlation table serves simply as a descriptive precursor to a more complex analysis (e.g., multiple regression, canonical correlation), then the correlation matrix should be presented with no references made to *p*-values either within or at the foot of the table.

Flaw 4: Over-reliance on Null Hypothesis Significance Tests of Correlation Coefficients

The literature is replete with calls for the reporting of effect sizes (e.g., Cohen, 1988; Daniel, 1998a, 1998b; Ernest & McLean, 1998; Knapp, 1998; Levin, 1998; McLean & Ernest, 1998; Nix & Jackson Barnette, 1998a, 1998b; Thompson, 1996, 1998a, 1998b, 1999). Even the strongest proponents of NHSTs concur that statistically significant results should be accompanied by one or more measures of practical significance (Barnette & McClean, 1999). Nevertheless, relatively few researchers consistently report estimates of effect size. According to Thompson (1998a), many researchers appear to be under the delusion that *p*-values (a) test result importance, (b) test result replicability, and (c) evaluate effect magnitude.

Even though correlation coefficients can be converted to effect size estimates with relative ease, a paucity of analysts do so. Yet, nowhere is it clearer that the test statistic underlying the NHST is largely dependent on the sample size than is the case for the product-moment correlation coefficient. A test of the hypothesis concerning a population correlation coefficient, like that for all other families of the general linear model, takes the general form:

$$\frac{\textit{Proportion of variance explained}}{\textit{Proportion of variance unexplained}}$$

Utilizing the fact that R^2 is the proportion of variance explained for a SLR model and 1 - R^2 is the proportion of variance unexplained, the test statistic for Pearson *r* generalizes to

$$t = \sqrt{\frac{R^2}{\frac{1 - R^2}{n - 2}}} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

which has a *t* sampling distribution for samples selected from a normal population. A close examination of the numerator of the right hand side of the above equation indicates that, holding the correlation constant, as the sample size (*n*) increases so does the value of *t*, and, consequently, the probability of rejecting the null hypothesis. Indeed, as derived by Pearson and Hartley (1962) and re-emphasized by Daniel (1998a), using $\alpha = .05$, whereas for a sample size of 3 the correlation coefficient has to be as large as .997 to be statistically significant, the correlation coefficient can be as low as .196 for a sample size of 100, .088 for a sample size of 500, .062 for a sample size of 1,000, and .020 for a sample size of 10,000.

Thus, alongside *p*-values, the practical significance (r^2)should be reported–for example, Cohen's (1988) criteria of .1 for a small correlational effect, .3 for a moderate correlational effect, and .5 for a large correlational effect. Reporting effect sizes should lead to the elimination of inappropriate language such as "highly significant" and "approaching significance," as well as result in the regular use of the phrase "statistically significant" (Carver, 1993; Cohen, 1994; Daniel, 1988, 1998a; Shaver, 1993; Thompson, 1996). Very recently, the American Psychological Association (APA) Board of Scientific Affairs (1999), who convened a committee called the Task Force on

Statistical Inference (TFSI), recommended in no uncertain terms that effect size estimates *always* be reported when reporting *p*-values. Hopefully, their recommendations will turn the tide.

In addition to the reporting of effect sizes, "what if" analyses could be reported. These analyses indicate how many more subjects are needed to obtain a statistical significance for the given correlation coefficient in cases in which the null hypothesis is not rejected, and how few cases are needed before a statistically significant relationship is no longer statistically significant (Daniel, 1998a). Furthermore, the confidence intervals for product-moment correlation coefficients could be reported (see Onwuegbuzie, 1999b for an example of reporting confidence intervals in a correlation matrix). Since the sampling distribution of the sample correlation coefficient for all values of the population correlation coefficient other than 0 is skewed, the sample correlation coefficient must be transformed in such a way that it has a sampling distribution which is approximately normal. Perhaps the most popular transformation is *Fisher's Z* transformation. This transformation statistic is defined as

$$|Z| = 0.5 \log_e \left(\frac{1 + |r|}{1 - |r|} \right)$$

where $\log_e$ is the natural logarithm and the "| |" indicates that the number contained in it can be either positive or negative. For example, the Fisher *Z*-value which corresponds to r = 0.828 is

$$Z = 0.5 \log_e \left(\frac{1 + 0.828}{1 - 0.828} \right) = 0.5 \log_e \left(\frac{1.828}{0.172} \right)$$

$$= 0.5 \log_e (10.6279) = 0.5(2.3635) = 1.182.$$

A simple way of obtaining the Fisher *Z*-values is to use the tables that are provided in many standard statistics textbooks. Such tables give the value of *Z* for values of r from 0 to 1.00. (If *r* is negative, the *Z* value obtained becomes negative). If the exact value of *r* is not listed, interpolation is used to obtain the corresponding *Z*-value. Conveniently, the distribution of *Z* is approximately normal regardless of the size of *n*, with a mean Z_{pop}, which corresponds to ρ, and a standard deviation given by

$$\sigma_z = \frac{1}{\sqrt{n - 3}}$$

A (1 - α)% confidence interval for Z_{pop} is

$$Z \pm (Z_{\alpha/2}) \cdot \sigma_z$$

Thus, the procedure for constructing 95% confidence intervals is as follows:

1. Transform *r* to Fisher *Z* using the equation above or the Fisher's *Z* transformation table.
2. Compute the standard error of *Z*.
3. Find a (1 - α)% confidence interval for Z_{pop}
4. Use the Fisher's *Z* table to transform the lower and upper confidence limits for

Z_{pop} back to *r* values.

In addition to allowing the researcher to test nil null hypotheses, confidence intervals also provide an approximate index of statistical power, with narrow intervals indicating high power and wide intervals indicating low power (Cohen, 1994). Unfortunately, currently, very few researchers report confidence intervals for correlation coefficients, probably because the major statistical packages do not provide this information. Thus, we hope that future versions of statistical software will allow this analysis to be performed.

Flaw 5: Conducting Tests of Statistical Significance for Reliability and Validity Coefficients

Because estimates of reliability and validity are in the form of correlation coefficients, many researchers undertake NHSTs of these coefficients using the nil null hypothesis (Huck & Cormier, 1999). However, as Thompson (e.g., Thompson, 1994b, 1996, 1998, 1999) and Daniel and his colleagues (Daniel, 1998a; Witta & Daniel, 1998) have argued, such tests are inappropriate. This is because large reliability and validity coefficients typically are statistically significant even when the sample sizes that underly them are small (Thompson, 1994b), whereas small coefficients will eventually become statistically significant as the sample size is increased (Huck & Cormier, 1999), due to the influence of sample size on NHSTs of correlation coefficients discussed above.

Moreover, reliability and validity coefficients are sample specific, and thus, statistically significant coefficients are neither necessarily replicable not generalizable (Witta & Daniel, 1998). Thus, rather than utilizing NHSTs of reliability and validity

coefficients, effect sizes should be used to assess the adequacy of instrument scores generated with specific samples. For example, Nunnally's (1978) criteria could be used for assessing the reliability of scores on non-cognitive measures for a specific sample, namely, that estimates of .70 and above are deemed to be adequate. For scores on measures of cognitive performance, .80 often is used as the cut-off point (e.g., Sattler, 1990), although some (e.g., Gay, 1999) recommend that .9 be used. Conversely, expressing a more liberal (though less rigid) view, Pedhazur and Schmelkin (1991, p. 110) noted that specific cutoffs should be avoided in favor of the researcher's judgment as to the "amount of error he or she is willing to tolerate given the specific circumstances of the study."

Flaw 6: Correcting for Attenuation

As stated earlier, one of the assumptions underlying a NHST of statistical significance is that both variables involved are measured without error (Myers, 1986). Obviously, this is seldom the case when dealing with social, behavioral, psychological, and educational variables. Unfortunately, when measurement errors are present, the relationship computed form the sample data will systematically underestimate the strength of the association in the population (Huck & Cormier, 1999). That is, errors of measurement produce biased estimates of the correlation coefficient that attenuate the true relationship. The greater the measurement error, the more the correlation coefficient is attenuated.

As a result, as has been previously noted, the statistical significance of a correlation coefficient is also a function of the reliability of the scores generated by the

sample. In other words, reliability coefficients affect statistical power. Specifically, low reliability coefficients tend to lower statistical power. In cases when the null hypothesis is not rejected, and one or more of the measures generate scores that have a low or even moderate reliability coefficient, one cannot be certain whether the statistically nonsignificant result suggests viability of the null hypothesis or merely represents a statistical artifact.

Thus, some researchers who have knowledge of their sample-specific reliability coefficients adjust their correlation coefficients to account for the estimated amount of unreliability. These analysts use a correction-for-attenuation formula which yields an adjusted/disattenuated correlation coefficient that is *always* higher than the uncorrected, raw *r* (Huck & Cormier, 1999). However, correcting for attenuation is an extremely controversial technique because it is subject to misapplication and misinterpretation (Muchinsky, 1996). For example, some researchers incorrectly claim that the method of correcting for attenuation improves the predictive accuracy of measures.

A common misapplication stems from the practice by many meta-analysts of disattenuating findings from individual studies before aggregating them into a composite score (i.e., effect size measure) . Unfortunately, because researchers are inconsistent in the statistics that they use to estimate reliability (i.e., internal consistency, test-retest, and parallel forms), these meta-analysts end up violating a major assumption of classical measurement theory that invalidates the interchangeability of different types of reliability (Cronbach, 1947). In other words,

aggregating findings that have been disattenuated using different measures of reliability seriously affects the validity of the resultant effect size estimates. Moreover, because the majority of researchers presently do not report sample-specific reliability coefficients in their reports, meta-analysts who prefer to disattenuate the findings of original researchers are left with incomplete data. Whether the analyst removes studies that do not present reliability coefficients from the analysis or imputes values (however determined) for missing reliability coefficients, there is no doubt that the composite effect size estimates will be biased.

Another area of controversy surrounding correcting for attenuation centers around which formula to utilize. Some theorists believe that Spearman's (1910) double correction formula should be used, namely:

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

where ρ_{xy} is the corrected validity coefficient, r_{xy} is the obtained sample correlation coefficient, r_{xx} is the reliability of scores yielded by the measure of the independent variable, and r_{yy} is the reliability of scores generated by the measure of the dependent variable. This formula corrects for unreliability of scores generated by measures of both the independent and dependent variables.

Other measurement theorists advocate the single correction formula

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

or

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

where the first single correction formula corrects for unreliability in scores generated by measures of the dependent variable only, and the second single correction formula adjusts for unreliability in scores yielded by measures of the independent variable only. These corrections can be useful in cases in which the reliability of scores on one of the variables is unknown to the researcher (e.g., when using data from standardized achievement tests in which only total scores are reported by the agency administering the test). Of the two single correction formulae, the former is utilized more often than is the latter (Muchinsky, 1996).

Also disputed is to which type of reliability coefficient the correction formulae should be applied. Historically, some theorists (e.g., Johnson, 1950) have noted that test-retest reliability coefficients should be used in correction formulas, whereas some (e.g., Guilford, 1954) have advocated that reliability coefficients of equivalence should be employed, and whereas others (e.g., Nunnally, 1978) have advanced internal consistency estimates. The debate continues today.

Whichever correction formula is used and whatever reliability estimate is utilized, researchers should never report disattenuated correlation coefficients in isolation. When these coefficients are presented, so should the raw (unadjusted) correlation

coefficients. Displaying disattenuated correlation coefficients alongside their unadjusted counterparts will allow the reader to assess the impact of unreliability on each bivariate relationship. In addition, the authors should explain which correction formula(e) have been used.

Flaw 7: The Crud Factor and Positive Manifold

As demonstrated above, as the sample size increases, so does the probability of rejecting the null hypothesis of no relationship between two variables. Indeed, theoretically, given a large enough sample size, the null hypothesis always will be rejected (Cohen, 1994). Hence, it can be argued that "everything correlates to some extent with everything else" (Meehl, 1990, p. 204). Meehl referred to this tendency to reject null hypotheses in the face of trivial relationships as the *crud* factor.

In support of the contention of the existence of a crud factor, Standing, Sproule, and Khouzam (1991), who computed a 135 x 135 matrix of correlations using 2,058 cases, found that, on average, each variable correlated at a statistically significant level ($p < .05$) with 41% of the other variables, although the absolute magnitude of the correlations averaged only .07. This finding not only provides a compelling example of the danger of relying solely on NHSTs, but also of the importance of selecting relationships of interest carefully, preferably stemming within a sound theoretical framework. In a similar analysis utilizing an extremely large sample, Meehl and Lykken (as cited in Cohen, 1994) conducted a study of 57,000 high school students in which cross tabulations for 15 Minnesota Multiphasic Personality inventory (MMPI) items yielded 105 chi-square tests of association--all of which were statistically significant,

with 96% of them being statistically significant at $p < .000001$.

Similar to the crud factor is the statistical artifact called a "positive manifold," in which individuals who perform well on one ability or attitudinal measure tend to perform well on other measures in the same domain (Neisser, 1998). For example, Tucker, Bass, and Daniel (1992) measured 106 university professors and administrators on three variables tracing the outcomes of transformational leadership (satisfaction, effectiveness, and extra effort) as reflected in their respective subscale Multifactor Leadership Questionnaire subscale scores. Because these outcomes are theoretically related, it is not surprising that the intercorrelations among the three sets of scores were characterized by positive manifold–all three correlations exceeded .65, with a mean across the correlations of .71.

Flaw 8: Inferring Causation from Correlation Coefficients

In interpreting correlation coefficients, researchers often infer cause-and-effect relationships, even though such relationships can, at best, only be determined from experimental studies. Scientific experiments can frequently make a strong case for causality by carefully controlling the values of all variables which might be related to the ones under study. Then if the dependent variable is observed to change in a predictable way as the value of the independent variable changes, the most plausible explanation would be a causal relationship between the independent and the dependent variable. In the absence of such control and ability to manipulate the independent variable, we must admit the possibility that at least one more unidentified variable is influencing both the variables under investigation.

This does not mean that correlational analysis may never be used in drawing conclusions about causal relationships. A high correlation in many uncontrolled studies carried out in different settings *can provide support for causality*--as in the case for the relationship between cigarette smoking and lung cancer. That is, correlations can be used to rule in or to eliminate (under conditions of replication) the possibility of a causal relationship. Kenny (1979) distinguished between correlational and causal inferences, noting that four conditions must exist before a scientist may appropriately claim that *X* causes *Y*: (a) time precedence (*X* must precede *Y* in time), (b) functional relationship (*Y* should be conditionally distributed across *X*), (c) nonspuriousness (there must not be a third variable *Z* that causes both *X* and *Y*, such that when *Z* is controlled for, the relationship between *X* and *Y* vanishes, and (d) vitality (a logistical link between *X* and *Y* that substantiates the likelihood of a causal link (such as would be established via controlled experimental conditions). Taken together, these four conditions lead the researcher to *infer* causality: "The law of causation is a conceptual figment extracted from phenomena, it is not of their very essence" (Pearson, 1911, p. 157).

Hence, substantiating causal links in uncontrolled (correlational) studies is a very elusive and futile task. Thus, researchers should pay special attention when interpreting findings stemming from correlation coefficients. Unfortunately, some researchers and policy makers are prone to relatively loose interpretations of such findings.

Flaw 9: Inappropriate Use of Hotteling's t-Test When Comparing Correlated Correlation Coefficients

Situations arise in which a comparison of the magnitude of two correlation

coefficients is of interest. For example, we might be interested in determining whether the relationship between test anxiety and test performance for female high school students is different from that of their male counterparts. Because the sampling distribution of *r* is skewed, we do not compare the correlations directly, but compare their corresponding Fisher *Z*-values. The z-test for testing independent *r*'s is given by

$$z = \frac{Z_1 - Z_2}{\sigma_{z_1 - z_2}}$$

where

$$\sigma_{z_1 - z_2} = \sqrt{\sigma^2_{z_1} + \sigma^2_{z_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

The right hand side of the first equation is the difference between Z_1 (i.e., Fisher's *Z*-value for the correlation pertaining to the first sample) and Z_2 (i.e., Fisher's *Z*-value for the correlation pertaining to the second sample) divided by the standard error of the difference. A confidence interval for the difference between two independent population correlations, $\rho_1 - \rho_2$, is given by the *r* transformations of the lower and upper limit obtained from

$$(Z_1 - Z_2) \pm z_{(\alpha/2)} (\sigma_{z_1 - z_2})$$

In addition to conducting NHSTs of independent correlations, effect size measures should be reported. The most common effect size measure is the difference

between the two Fisher's Z-values (i.e., $Z_1 - Z_2$). Cohen's (1988) criteria for product-moment correlations (i.e., .1 = small, .3 = medium, and .5 = large) could be utilized to determine the magnitude of effect sizes. At present, very few researchers formally compare independent correlation coefficients, probably because the major statistical packages do not conduct such analyses. Thus, we hope that future versions of statistical software will allow these analyses to be undertaken.

If a single sample is drawn from one population, and one is interested in comparing two correlation coefficients that are computed on the basis of the sample data, the null hypothesis is that there is no difference between the two correlations in the single population associated with the study. For instance, one may be interested in comparing the relationship between self-esteem and achievement to that between anxiety and achievement for the same sample. In this case, two *correlated* correlations are being compared. The procedure for performing a NHST of two correlated correlations is different than that for comparing two independent correlations.

Presently, the most common technique for comparing correlations is Hotelling's *t*-test (Hotelling, 1940). Unfortunately, this technique has a serious flaw. Although Hotelling's *t*-test is exact, it only tests the nil null of equal correlation coefficients when the sample variance of both variables of interest equals the corresponding population variance, which very rarely occurs. In fact, Hotelling (1940, pp. 276-277) warned, "The advantages of exactness and of freedom from the somewhat special trivariate normal assumption are attained at the expense of sacrificing the precise applicability of the results to other sets of the predictors."

A more appropriate statistic for comparing two correlated coefficients with a common variable was proposed by Meng, Rosenthal, and Rubins (1992). Specifically, these authors derived the following *Z*-test for assessing the statistical significance of the difference between two sample correlation coefficients r_{yx1} and r_{yx1}, where variables X_1 and X_2 are predictors of the dependent variable *Y*:

$$Z = (z_{r1} - z_{r2}) \sqrt{\frac{N - 3}{2(1 - r_x)h}}$$

where *N* is the number of subjects, z_{r1} is the Fisher *Z*-transformation of the first correlation, z_{r2} is the Fisher *Z*-transformation of the second correlation, r_x is the correlation between the two predictor variables, X_1 and X_2,

$$h = \frac{1 - f\,\overline{r^2}}{1 - \overline{r^2}} = 1 + \frac{\overline{r^2}}{1 - \overline{r^2}}(1 - f)$$

$$f = \frac{1 - r_x}{2(1 - \overline{r^2})}$$

and r^{-2} is the mean of the two correlations involving the predictor variables. A 95% confidence interval for the difference in Fisher *z*'s is

$$z_{r1} - z_{r2} \pm 1.96\sqrt{\frac{2(1 - r_x)h}{N - 3}}$$

Again, we hope that future versions of statistical software will allow these analyses to

be performed.

Flaw 10: Failure to Assess the External and Internal Replicability of Correlation Coefficients

There is little doubt that external replication is the essence of science. Thompson (1994) provides four reasons for replication. First, an individual study cannot explain adequately a phenomenon without introducing bias. Second, the findings from an individual investigation are limited by design and measurement flaws. Third, an individual study is limited by analytical methods. Fourth, an individual study is affected by the inherent limitations of NHSTs.

Types of replication include direct replication, simultaneous replication, systematic replication, and clinical replication (Gay, 1999). Direct replication involves replication by the same investigator using the same or different subjects in a specific setting. Simultaneous replication refers to replications undertaken on subjects with the same characteristics at the same location and at the same time. Systematic replication pertains to replication which follows direct replication, involving different investigators, behaviors, or settings. Finally, clinical replication involves the development treatment programs, comprising two or more interventions which have been found to be effective individually (Gay, 1999).

Regardless of the type of external replication conducted, the more that results are replicated, the more confidence can be placed on the original finding. Indeed, it is only by replicating findings across different settings and using different samples that we can hope to form theoretical generalizations. Thus, researchers should not only

compare their correlation coefficients to those obtained in previous studies, but should also encourage and attempt external replications.

Unfortunately, as noted by Thompson, few social scientists conduct external replication analyses, probably due to time, resources, or energy constraints (Thompson, 1994). Thus, we recommend that internal replications be undertaken in which the stability of sample correlation coefficients are assessed using data from the available sample. Although internal replications are inferior to external replications, the former still provide useful information about the stability of coefficients, and thus should be utilized routinely--even when external replications are possible!

The three most common classes of internal replication utilize either cross-validation, jackknife, or bootstrap techniques. For the Pearson product-moment correlation coefficient, cross-validation involves dividing the sample into two approximately equally sized groups, computing a correlation coefficient for the first group, and then using the second group to attempt to confirm this coefficient. Jackknife techniques involve computing a series of correlation coefficients, with groups of subjects of an equal size (usually one at a time) being deleted from each analysis once only. Finally, bootstrap methods involve resampling the same dataset repeatedly (i.e., thousands of times), and then computing the correlation coefficient for each sample. The mean correlation coefficient is then computed and compared to the original correlation from the full sample in order to assess stability.

At present, very few social scientists conduct internal replications. Of those who do, these replications tend to occur after multivariate models have been fitted. Virtually

no researchers conduct internal replications to examine the stability of the Pearson product-moment correlation coefficient, despite the fact that (a) this coefficient is a member of the general linear model family, (b) this coefficient is the most widely used to conduct inferential analyses, and (c) this coefficient Is as susceptible to instability as are other members of the general linear model. Thus, what follows is an example using a small heuristic dataset to illustrate how internal replication techniques can identify unstable correlation coefficients derived from the full sample.

Heuristic Example

Recently, Onwuegbuzie, Slate, Paterson, Watson, and Schwartz (in press) conducted a study investigating correlates of achievement among students enrolled in several sections of a graduate-level quantitative-based educational research course at a southeastern university. The theoretical framework for this investigation, though not presented here, can be found by examining the original study. Although several independent variables were examined by Onwuegbuzie et al., we will restrict our attention to one of them, namely, anxiety resulting from fear of the statistics instructor.

Fear of the statistics instructor is one of the six subscales of the Statistical Anxiety Rating Scale (STARS; Cruise & Wilkins, 1980). This subscale is a 5-item, 5-point Likert-format instrument which assesses students' perceptions of their statistics instructor. Scores on this subscale range from 5 to 25, such that high scores represent high levels of anxiety induced by fear of the statistics teacher. According to Onwuegbuzie et al. (in press), scores pertaining to the *fear of the statistics instructor* subscale had a classical theory alpha reliability coefficient of .83. This represents an

acceptable level (cf., *Flaw 5*). Evidence of construct-related validity was obtained via correlations ranging from .50 to .75 between scores on the *fear of the statistics instructor* subscale and scores on the other five subscales of the STARS. For purposes of this heuristic example, the total score of the *fear of the statistics instructor* subscale (M = 12.27, SD = 4.04) was used as the independent variable.

The dependent variable chosen for this heuristic example was level of achievement in the educational research course, which was measured using students' course averages (M = 88.51, SD = 4.92). Students' course averages comprised evaluation of research articles, written research proposals, orally presented research proposals, and conceptual knowledge (as measured by five untimed in-class examinations). For the present study, the Statistical Package for the Social Sciences (SPSS Inc., 1999) was utilized to obtain descriptive statistics, as well as for the correlational and regression analysis. The Statistical Analysis System (SAS Institute Inc., 1999) was used to perform the jackknife analysis. Finally, Amos 4.0 (Arbuckle & Wothke, 1999) was utilized to undertake the bootstrap analysis.

Treating these two variables in isolation, the sample size of 121 reflected a statistical power of .92 (cf., *Flaw 3*) to detect a moderate bivariate relationship (i.e., r = .30) with a statistical significance level of .05 (Cohen, 1988). Thus, the level of power for this heuristic example was high (cf., *Flaw 3*), and indeed much higher than is the case for most NHSTs in educational research (see for example, Cohen, 1994, 1997).

An inspection of the scatterplot (cf., *Flaw 1*) suggested no evidence of a non-linear relationship between the two variables. In addition, an examination of the

histogram and the expected normal probability plot pertaining to the achievement variable suggested normality. Furthermore, the skewness and kurtosis coefficients indicated that the achievement variable was approximately normal. Specifically, both the skewness coefficient of -0.30 (*SE* = .22) and the kurtosis coefficient of -0.21 (*SE* = .44) were small enough for the distribution to be considered normal. Indeed, the *z*-values corresponding to both skewness (z = -1.35) and kurtosis (z = -0.48) coefficients were not significant ($p > .05$). Also, the Shapiro-Wilk test (Shapiro & Wilk, 1965; Shapiro, Wilk, & Chen, 1968) did not indicate that the distribution of educational research achievement scores was non-normal ($W = .98$, $p > .05$).

Similarly, the independent variable, *fear of the statistics instructor*, appeared to be normally distributed. Specifically, the histogram and the expected normal probability plot suggested normality. In addition, both the skewness coefficient of 0.43 (*SE* = .22) and the kurtosis coefficient of 0.35 (*SE* = .44) were small enough for the distribution to be considered normal. In fact, the *z*-values corresponding to both skewness (z = 1.95) and kurtosis (z = 0.80) coefficients were not statistically significant ($p > .05$). Also, the Shapiro-Wilk test did not indicate that the distribution of educational research achievement scores was non-normal ($W = .96$, $p > .05$).

The fact that the scores pertaining to both variables appeared to be normally distributed justified the use of the Pearson product-moment correlation coefficient (cf., *Flaw 1*) for examining the relationship between fear of the statistics instructor and achievement. The Pearson product-moment correlation coefficient indicated a correlation between these two variables of -.1761. (Four decimal places were used

rather than the usual two decimal places to accommodate the jackknife and bootstrap analyses.) The *p*-value associated with this correlation was .0533. Interestingly, some *p-value analysts* may inappropriately use terms such as "approaching significance" or "marginally significant" to describe this correlation. Indeed, if a *p*-value was the only criterion used to describe a Pearson's product-moment correlation coefficient, a *p*-value such as the above, may lead some analysts to reject the null hypothesis of a zero relationship in the population, and some researchers *not* to reject the null hypothesis--depending on how many decimal places (e.g., 2 vs. 4 decimal places) are used for the correlation coefficient and/or the *p*-value.

Although, strictly speaking, this correlation is not statistically significant at the 5% level, this example strongly reinforces the important point that analysts of empirical data should never rely merely on *p*-values to make statistical inferences about a sample (cf., *Flaw 4*). Indeed, use of Cohen's (1988) criteria for effect sizes to interpret the *educational/practical significance* of the present correlation renders the discussion about its statistical significance almost moot, because regardless of the level of statistical significance, the correlation is quite small.

The problem in deciding whether the correlation is statistically significant also bolsters our contention that confidence intervals for correlations also should be reported. For the present example, using Fisher's *Z*-transformation, a 95% confidence interval for the relationship between *fear of the statistics instructor* and achievement was -.3438 to .0025. Since this interval includes 0, we would not reject the null hypothesis at the 5% level--which supports our earlier decision. Moreover, the

confidence interval is relatively wide, ranging from a zero effect to a moderate effect. This suggests that the correlation coefficient may not be very stable--a finding that would not have emerged if a confidence interval had not been constructed.

Influence Diagnostics

A simple linear regression (SLR) model was fitted, using SAS, with anxiety (i.e., fear of the statistics instructor) as the independent variable and achievement as the dependent variable. This analysis yielded the following estimates of the model parameters (the standard errors of the coefficients are in parentheses):

achievement = 90.592 -0.213*Anxiety

(1.407) (0.109)

This SLR model allowed us to check further the adequacy of the correlation coefficient. In particular, an inspection of the studentized residuals generated from the model (Myers, 1986) suggested that the assumptions of normality, linearity, and homoscedasticity were met. Using the Bonferroni adjustment, none of the studentized residuals suggested that outliers were present. Specifically, only 6 of the 121 studentized residuals had absolute values larger than 2.0, with 5 of these values being less than 2.5, and the remaining value being 2.73. Thus, the studentized residuals did not give any major cause for concern.

Additionally, the following influence diagnostics were examined: (1) the number of estimated standard errors (for each regression coefficient) that the coefficient changes if the *i*th observation were set aside (i.e., DFBETAS); (2) the number of estimated standard errors that the predicted value changes if the *i*th point is removed

from the data set (i.e., DFFITS); (3) the reduction in the estimated generalized variance of the coefficient over what would have been produced without the *i*th data point (i.e., COVRATIO), and (4) a measure of standardized distance from the *i*th point of the anxiety variable to the data center in the variable (i.e., HAT diagonal). Using criteria recommended in the literature (e.g., Myers, 1986; Sen & Srivastava, 1990), it was revealed that three observations were potential outliers, namely, subjects 73, 96, and 114. Subject 73 had a relatively large HAT diagonal, suggesting strong leveragability (i.e., disproportional influence on the anxiety coefficient). This subject also had relatively large DFFIT and DFBETA values for the intercept and slope. The DFFIT value of -0.32 indicates a one-third of a standard error decrease in achievement due to the inclusion of this subject. The DFBETA value of -0.29 on the anxiety coefficient suggests that without the 73rd participant, the regression coefficient decreases by 0.29.

Subject 96 had a DFFIT value of -0.34, also indicating a one-third of a standard error decrease in achievement due to her/his inclusion. Also, the DFBETA value of 0.29 on the anxiety coefficient suggests that the exclusion of this participant would increase the anxiety coefficient by 0.29. Finally, subject 114 had an absolute studentized residual that was greater than 2, with a DFFIT value which suggested a one-third of a standard error decrease in achievement, and a DFBETA value of .27.

Bootstrap Analysis

The second method of assessing result replicability utilized the bootstrap method, which was developed by Efron and his colleagues (Diaconis & Efron, 1983; Efron, 1979; Efron & Tibshirani, 1993). Bootstrap analyses involve resampling the

same dataset (i.e., sampling with replacement) a specified large number (typically thousands) of times and computing the statistics of interest for each sample. These statistics are then averaged, and the standard deviation of the bootstrap-estimated sampling distribution (i.e., standard error of estimate) is derived. The standard deviation of the sampling distribution provides an estimate of the variability of the sample statistics given fluctuations in the sample. In order to justify making inferences on the bootstrap estimates, thousands of resamples are required.

For the purpose of this heuristic example, 10,000 resamples were undertaken on the SLR model. The mean regression coefficient for the anxiety variable was -0.213. Encouragingly, this value is identical to the regression coefficient for the original sample. In addition, the standard deviation of the 10,000 bootstrap estimates (i.e., standard error of estimate) was .001, suggesting strongly that the regression coefficient and, consequently, the correlation coefficient were both extremely stable. Because the standard error of estimate was miniscule, the 95% confidence interval (not reported here) was extremely narrow.

Jackknife Analysis

The third method of assessing result replicability utilized the jackknife method (Crask & Perreault, 1977) (cf., *Flaw 10*). This procedure entailed conducting 121 separate correlations (each examining the same relationship between anxiety and achievement), wherein each analysis involved dropping the *i*th participant until *every* subject had been eliminated exactly *once*. That is, each of the resultant 121 correlations utilized 120 participants (i.e., n-1 participants, where n = the total sample

size). The 121 *r* values which were generated from these models were examined for stability. The summary statistics pertaining to this analysis are presented in Table 1.

Insert Table 1 about here

As can be seen in Table 1, encouragingly, the standard deviation (.0074) about the mean jackknife correlation estimate (.1761) was extremely small. Also, assuming that the sample estimates of the correlation coefficient are normally distributed (as suggested by the closeness of the mean and median values for the correlation estimates), it can be seen that the 95% confidence interval about the parameter estimate lies between -.1775 and -.1748. Encouragingly, this interval is not only very narrow, but it contains the estimate calculated using the complete data (i.e., $r = -.1764$). This finding of a stable correlation coefficient echoes the result from the bootstrap analysis.

However, caution should be exercised in interpreting this interval because, although the skewness coefficient (.08) pertaining to the jackknife correlation estimates was small relative to its standard error (.22), the kurtosis coefficient (3.06) was extremely large as compared to its standard error. Indeed, the *z*-value associated with the kurtosis coefficient was 7.00. Figure 1, which presents the histogram of the 121 jackknife correlation coefficients, supports this finding of a small skewness coefficient and a large kurtosis coefficient. Moreover, from this graph, it can be seen that the participants whose removal from the sample generated a correlation coefficient of -

.2019 (subject 96), -.2025 (subject 114), and -.1475 (subject 73) may be outliers. These three subjects also were highlighted via the influence diagnostic analysis. This illustrates another useful role of jackknife analyses--identifying potential outliers (cf., *Flaw 1*). Interestingly, if these three subjects are removed from the sample, the correlation increases to -.2010 ($p < .05$), changing our conclusion about the correlation coefficient from statistically nonsignificant to statistically significant, although the effect size is still relatively small.

Insert Figure 1 about here

The final step of the jackknife analysis was to examine the *p*-values associated with each jackknife correlation estimate. Table 2 presents the descriptive statistics pertaining to the jackknife *p*-values. (Ninety-five percent confidence intervals were not constructed due to the large skewness and kurtosis coefficients.) It can be seen that the *p*-values ranged from .0270 to .1069. The mean *p*-value was .0545, reflecting the dilemma discussed above regarding the level of statistical significance of the bivariate correlation. Even more disturbing was the fact that, of the 121 *p*-values, 33 (27.3%) would have been declared statistically significant at the 5% level, whereas 88 (72.7%) would not have been declared statistically significant. In other words, if only one member of the sample had not been included in the study, relying only on *p*-values to interpret the "significance" of the results would have led to inconsistent conclusions, depending on which participant was absent. This undoubtedly provides the most

compelling example of the importance of routinely interpreting effect sizes alongside *p*-values (cf., *Flaw 4*).

Insert Table 2 about here

The fact that using the full sample would have led to a statistically nonsignificant correlation coefficient, but that removing only one participant from the study yielded statistical significance more than one-fourth of the time, indicates that the internal replication error rate, to which we will refer as the *Type V* error rate, far exceeded the nominal Type I error rate (i.e., 5%). Thus, jackknife analyses are extremely useful in providing information about Type V error rates.

Conclusions

The purpose of the present paper was to provide an in-depth critical analysis of the use of correlation coefficients. We argue that many social scientists do not exhibit the same care and consideration when calculating correlation coefficients as they do in conducting more complex analyses of the general linear model. Simply put, many analysts take the correlation coefficient in general and Pearson's product-moment correlation coefficient in particular for granted.

Ten flaws are identified and discussed which are made by many researchers when examining bivariate relationships. From these flaws, the following "ten commandments" are appropriate when utilizing Pearson product-moment correlation coefficients:

(1) Always check statistical assumptions *prior* to using Pearson's *r*, as well as after the correlation has been computed.

(2) Always adjust for Type I error when conducting multiple NHSTs of correlations.

(3) Always be cognizant of the power of NHSTs of correlations, preferably before the data collection stage, and, at the very least, at the data analysis stage.

(4) When making inferences about the Pearson *r* value, always interpret effect sizes.

(5) Do not conduct tests of statistical significance for reliability and validity coefficients.

(6) Do not report disattenuated correlation coefficients without also presenting the raw coefficients.

(7) Do not correlate variables without a theoretical framework.

(8) Avoid inferring causation from a correlation coefficient, regardless of how large the effect size is.

(9) Do not use Hotelling's *t*-test when comparing correlated correlation coefficients

(10) Conduct external replications when possible, and, in their absence, always undertake internal replications.

A heuristic example was provided to illustrate how jackknife and bootstrap methods can assess the stability of product-moment correlation coefficients derived

from the full sample. It was demonstrated that, even when correlation coefficients are stable, as in the present example, large internal replication errors (i.e., Type V errors) may prevail. As such, we hope that this paper will be useful for both beginning and experienced researchers. Moreover, we hope that our efforts will help to motivate others to pay more attention to our most commonly-used inferential statistic.

References

American Psychological Association Board of Scientific Affairs. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Arbuckle, J.L., & Wothke, W. (1999). *Amos 4.0 Users' guide.* Chicago, IL: Smallwaters Cooperation.

Barnette, J.J., & McClean, J.E. (1999, November). *Empirically based criteria for determining meaningful effect size.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, Alabama.

Carver, R.P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287-292.

Coblick, G.E., Halpin, G., & Halpin, G. (1998, November). *The development of statistical inference.* Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, LA.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Psychology, 65*, 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-443.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: John Wiley.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Cohen, J. (1997). The earth is round (p < .05). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (eds.), *What if there were no significance tests?* (pp. 21-36) Mahwah, New Jersey: Lawrence Erlbaum Associates.

Crask, M.R., & Perreault, W.D. (1977). Validation of discriminant analysis in marketing research. *Journal of Marketing Research, 14,* 60-68.

Cronbach, L.J. (1947). Test "reliability": Its meaning and determination. *Psychometrika, 12*, 1-16.

Cruise, R.J., & Wilkins, E.M. (1980). *STARS: Statistical Anxiety Rating Scale*. Unpublished manuscript, Andrews University, Berrien Springs, MI.

Daniel, L.G. (1988). [Review of Conducting educational research (3rd ed.)]. *Educational and Psychological Measurement, 48*, 848-851.

Daniel, L.G. (1989, January). *Use of the jackknife statistic to establish the external validity of discriminant analysis results*. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 305 382)

Daniel, L.G. (1997). Kerlinger's research myths: An overview with implications for educational researchers. Journal of Experimental Education, 65, 101-112.

Daniel, L.G. (1998a). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for editorial policies of educational journals. *Research in the Schools, 5*, 23-32.

Daniel, L.G. (1998b). The statistical significance controversy is definitely not over: A rejoinder to responses by Thompson, Knapp, and Levin. *Research in the Schools, 5*, 63-65.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American, 248*(5), 116-130.

Eason, S.H., & Daniel, L.G. (1989, January). Trends and methodological practices in several cohorts of dissertations. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 306 299)

Efron, B. (1979).Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Ernest, J.M., & McLean, J.E. (1998). Fight the good fight: A response to Thompson, Knapp, and Levin. *Research in the Schools, 5*, 59-62.

Fish, L.J. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development, 21*(3), 130-137.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Gay, L.R. (1999). *Educational research: Competencies for analysis and application* (6th ed.). New York: Merrill.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989).

The empire of chance. Cambridge: Cambridge University Press.

Guilford, J.P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics, 11*, 271-283.

Huberty, C.J., & Wisenbaker, J.M. (1992). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), *Advances in social science methodology* (Vol 2, pp. 169-208). Greenwich, CT: JAI Press.

Huck, S.W., & Cormier, W.H. (1999). *Reading statistics and research* (4th ed.). New York: HarperCollins College Publishers.

Johnson, H.G. (1950). Test reliability and correction for attenuation. *Psychometrika, 15*, 115-119.

Kenny, D. A. (1979). *Correlation and causality*. New York: John Wiley & Sons.

Kerlinger, F.N. (1960). The mythology of educational research: The methods approach. *School and Society, 88*, 149-151.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410-416.

Knapp, T.R. (1998). Comments on the statistical significance testing articles.

Research in the Schools, 5, 39-42.

Levin, J.R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*, 43-54.

Locke, L.F., Spirduso, W.W., & Silverman, S.J. (1987). *Proposals that work: A guide for planning dissertations and grant proposals* (2nd ed.). Newbury Park, CA: Sage.

Maxwell, S.E., & Delaney, H.D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth Publishing Company.

McLean, J.E., & Ernest, J.M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*, 15-22.

Meehl, P. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Meier, S.T., & Davis, S.R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. *Journal of Counseling Psychology, 37*, 113-115.

Meng, X-L., Rosenthal, R., & Rubin, D.B. (1992) Comparing correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Muchinsky, P.M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*, 63-75.

Myers, R.H. (1986). *Classical and modern regression with applications*. Boston, MA: Duxbury Press.

Nix, T.W., & Barnette, J. (1998a). The data analysis dilemma: Ban or abandon. A

review of null hypothesis significance testing. *Research in the Schools, 5*, 3-14.

Neisser, U. (1998). Rising test scores. In U. Neisser (Ed.), *The rising curve* (pp. 3-22). Washington, DC: American Psychological Association.

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Onwuegbuzie, A.J. (1999a, September). *Common analytical and interpretational errors in educational research*. Paper presented at the annual meeting of the European Educational Research Association (EERA), Lahti, Finland.

Onwuegbuzie, A.J. (1999b). Writing apprehension among graduate students: Its relationship to self-perceptions. *Psychological Reports, 84,* 1034-1039.

Onwuegbuzie, A.J. (1999c). *Correlates of achievement in graduate-level educational research courses*. Unpublished manuscript, Valdosta State University, Valdosta, Georgia.

Onwuegbuzie, A.J., Slate, J., Paterson, F., Watson, M., & Schwartz, R. (in press). Factors associated with underachievement in educational research courses. *Research in the Schools.*

Pearson, K. (1911). *The grammar of science* (3rd ed.). London: Adam and Charles Black.

Pearson, E.S., & Hartley, H.O. (Eds.). (1962). *Biometrika tables for statisticians* (2nd ed.). Cambridge, MA: Cambridge University Press.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

SAS Institute Inc. (1990). *SAS/STAT User's Guide* (Version 6.12) [Computer

software]. Cary, NC: SAS Institute Inc.

Sattler, J.M. (1990). *Assessment of children*. San Diego, CA: Author.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115-129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Sen, A.K., & Srivastava, M. (1990). *Regression analysis: Theory, methods, and applications*. New York: Springer-Verlag.

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test, for normality and complete samples. *Biometrika, 52,* 592-611.

Shapiro, S.S., Wilk, M.B., & Chen, H.J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association, 63,* 1343-1372.

Shaver, J. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61*, 293-316.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

SPSS Inc. (1999). *SPSS 9.0 for Windows*. [Computer software]. Chicago, IL: SPSS Inc.

Standing, L., Sproule, R., & Khouzam, N. (1991). Empirical statistics: IV. Illustrating Meehl's sixth law of soft psychology: Everything correlates with everything. *Psychological Reports, 69*, 123-126.

Stevens, J. (1996). *Applied multivariate for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper & Row.

Thompson, B. (1994a). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality, 62*, 157-176.

Thompson, B. (1994b). *Guidelines for authors. Educational and Psychological Measurement, 54*, 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998a, April). *Five methodological errors in educational research: The pantheon of statistical significance and other faux pas*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Thompson, B. (1998b). Statistical testing and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*, 33-38.

Thompson, B. (1999, April). *Common methodological mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Montreal, Canada. Retrieved October 11, 1999 from the World Wide Web: **http://acs.tamu.edu/~bbt6147/aeraad99.htm.**

Tucker, M.L., Bass, B.M., & Daniel, L.G. (1992). Transformational leadership's

impact on higher education satisfaction, effectiveness, and extra effort. In K.E. Clark, M.B. Clark, & D. P. Campbell (Eds.), *Impact of Leadership*. Greensboro, NC: Center for Creative Leadership.

Wilcox, R.R. (1997). *Introduction to robust estimates and hypothesis testing*. San Diego, CA: Academic Press.

Willson, V.L. (1980). Research techniques in *AERJ* articles: 1969 to 1978. *Educational Researcher, 9*(6), 5-10.

Witta, E.L., & Daniel, L.G. (1998, April). *The reliability and validity of test scores: Are editorial policy changes reflected in journal articles?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Table 1

Descriptive Statistics and 95% Percent Confidence Intervals about the Jackknife Estimates of Pearson's Product-Moment Correlation Coefficient (*r*) (Using 121 Resamples Each of Sample Size 120)

Summary Statistic	*r*
Mean	.1761
Median	.1764
Standard Deviation	.0074
Minimum	.1479
Maximum	.2019
Range	.0540
Skewness	.075
Standard Error of Skewness	.220
Kurtosis	3.059
Standard Error of Kurtosis	.437
95% Lower Bound	.1748
95% Upper Bound	.1775
Full Sample (n = 121)	.1761

Table 2

Descriptive Statistics for the Jackknife *p*-Values

Summary Statistic	*p*-Value
Mean	.0551
Median	.0539
Standard Deviation	.0106
Minimum	.0270
Maximum	.1069
Range	.0799
Skewness	1.077
Standard Error of Skewness	.220
Kurtosis	4.860
Standard Error of Kurtosis	.437
Full Sample (n = 121)	.0533

Figure 1: Histogram of 121 Jackknife Correlation Coefficients